IDENTIFYING ISOFORM SWITCHES IN BREAST CANCER

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ŞEVKI ONUR HENDEN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEP 2021

Approval of the thesis:

IDENTIFYING ISOFORM SWITCHES IN BREAST CANCER

submitted by **ŞEVKI ONUR HENDEN** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil KALIPÇILAR
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit OĞUZTÜZÜN
Head of Department, **Computer Engineering** _____

Prof. Dr. Tolga CAN
Supervisor, **Computer Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Volkan ATALAY
Computer Engineering, METU _____

Prof. Dr. Tolga CAN
Computer Engineering, METU _____

Assist. Prof. Dr. Ercüment ÇİÇEK
Computer Engineering, Bilkent University _____

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Şevki Onur HENDEN

Signature :

ABSTRACT

IDENTIFYING ISOFORM SWITCHES IN BREAST CANCER

HENDEN, Şevki Onur
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. Tolga CAN

Sep 2021, 44 pages

Characterizing the human genome's molecular functions and their variations across people is vital for understanding the cellular processes behind human genetic characteristics and diseases. With the advent of single-cell RNA sequencing (scRNA-seq), it is now possible to investigate gene expression in individual cells. Although a number of scRNA-seq bioinformatics tools are now available, many of them focus on overall gene expression levels and, as a result, often ignore heterogeneity caused by individual transcript expression. Differences in the relative abundance of expressed isoforms, such as those that occur between normal and diseased states, may have dramatic effects on phenotype or prognosis. This variation in expression may aid in the discovery of novel therapies as well as the better management of patients in certain situations. We propose a computational workflow for scRNA-seq data that identifies differential transcript usage from transcript abundances produced by widely used alignment tools such as Salmon. This approach enabled us to detect alterations in gene expression that were previously overlooked, in patients with breast cancer.

Keywords: breast cancer, isoform switch, manova, data analysis

ÖZ

MEME KANSERİNDE İZOFORM DEĞİŞİKLİKLERİNİN TANIMLANMASI

HENDEN, Şevki Onur
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Tolga CAN

2021 , 44 sayfa

İnsan genomunun moleküler işlevlerini ve varyasyonlarını karakterize etmek, insan genetik özelliklerinin ve hastalıklarının arkasındaki hücresel süreçleri anlamak için hayati önem taşır. Tek hücreli RNA dizilemesinin ortaya çıkmasıyla birlikte, gen ekspresyonunu hücreler özelinde araştırmak mümkün hale gelmiştir. Şu anda birçok scRNA-seq biyoenformatik analiz metodu mevcut olmasına rağmen, bunların çoğu genel gen ekspresyon seviyelerine odaklanır ve sonuç olarak, bireysel transkript ekspresyonunun neden olduğu heterojenliği göz ardı ederler. Normal ve hastalıklı durumlar arasında görece farklı sayılarda eksprese edilen izoformlar, fenotip veya prognoz üzerinde dramatik etkilere sahip olabilir. İzoformlardaki bu çeşitliliğin bulunması, yeni tedavilerin keşfedilmesine ve belirli durumlarda hastaların daha iyi yönetilmesine yardımcı olabilir. Salmon gibi hizalama araçları tarafından, scRNA-seq verileri için üretilen transkript ekspresyonundan, diferansiyel transkript kullanımını belirleyen bir analiz yöntemi öneriyoruz. Bu yaklaşım, meme kanserli hastalarda daha önce gözden kaçan gen ekspresyonundaki değişiklikleri tespit etmemize olanak sağladı.

Anahtar Kelimeler: meme kanseri, izoform değişikliği, manova, veri analizi

To my family

ACKNOWLEDGMENTS

I'd want to express my deepest gratitude to my supervisor, Prof. Dr. Tolga CAN, who has always mentored and encouraged me. His tremendous knowledge really aided me through the most difficult circumstances. It has been a pleasure to be guided by him, and I hope to work with him again in the future.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid
mRNA	Messenger RNA
scRNA-seq	Single cell RNA sequencing
RNA-seq	RNA sequencing
GEO	Gene Expression Omnibus
GTEX	Genotype-Tissue Expression
APA	Alternative polyadenylation
AS	Alternative splicing
DE	Differentially expressed
DTU	Differential transcript usage
FDR	False discovery rate
UTR	Untranslated region
NMD	Nonsense- mediated RNA decay
TPM	Transcript per million
MANOVA	Multivariate analysis of variance
ANOVA	Analysis of variance
PC	Principle component

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Definition

Due to alternative splicing, a gene may generate various mRNA sequences that code for functionally distinct protein isoforms. These changes may have a significant functional impact on organisms which may be both unicellular and multicellular. Single-cell sequencing methods have enabled researchers to individually characterize all cell types in a complicated population, such as a tumor or growing organ. Studies aided in the discovery of novel types of cells, thus improving our knowledge of both normal and disease-related biological processes and paving the way for the development of innovative therapeutic methods. Although many methods for performing scRNA-seq analysis are now available, many of them only assess overall gene expression levels and therefore do not include heterogeneity in transcript expression, especially for genes that exhibit multiple isoforms. Unbiased transcriptome analysis of control and treatment groups can help identify markers that alter functional modules and helps researchers understand the disease mechanisms.

Here is the outline for gene CLK1 in human shown in Figure 1.1 Each line represents a transcript and rectangles represents coding parts of the gene and between rectangles not coding parts. Most of the current research focuses on gene level differential expression where read counts created by aggregating the transcript level expressions. Also, extended this analogy to transcript level reads, Having one transcript to be significantly expressed , in other words , transcript level differential expression is not the focus of this thesis. Our focus to find behaviour change in pairs of isoforms in different conditions settings. This kind of information cannot be retrieved using con-

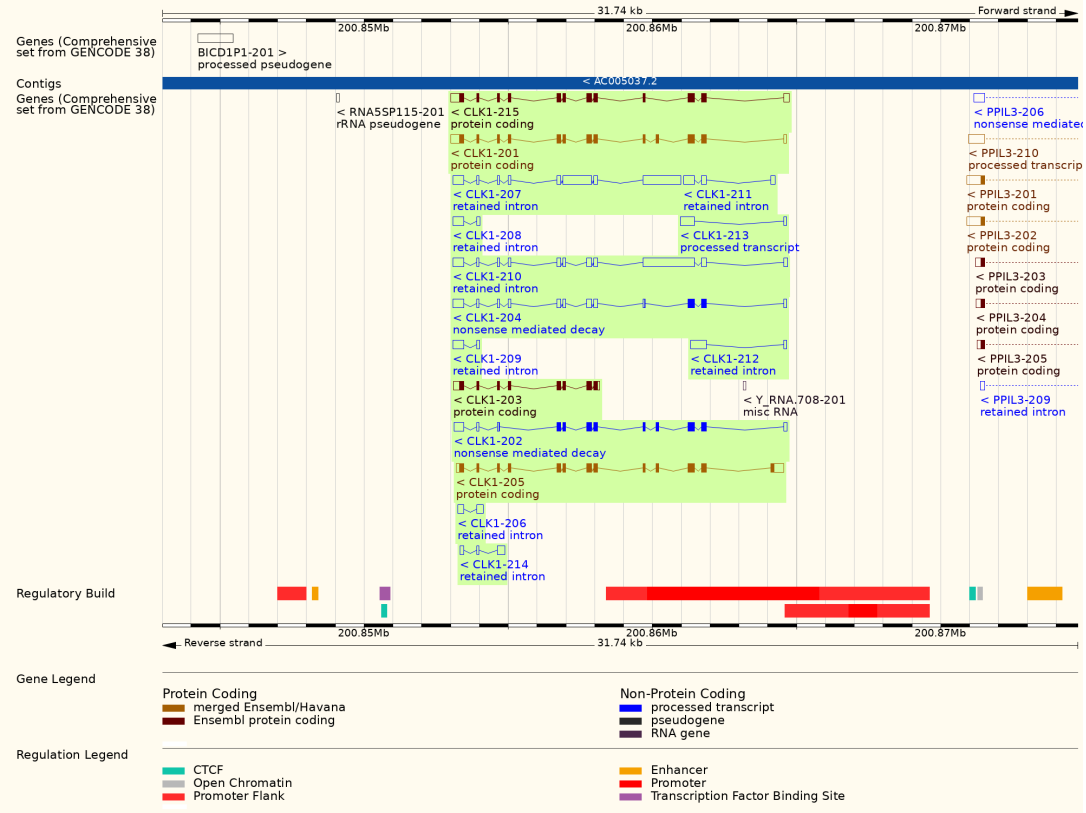

Figure 1.1: ENSEMBL entry for CLK1 gene

ventional differential expression methods. Significant expression of a transcript does not mean isoform switch or differential transcript usage. It takes two transcript to be in behavior change to be accepted as isoform switch. This kind of solution, will help scientists to uncover in certain conditions(tumor or disease) which parts of the gene is selectively transcribed and which parts are not.

1.2 Proposed Methods and Models

As a consequence of technological noise and dropouts, read counts for single cell RNA sequencing are often zero or near-zero. This is due to the restricted amount of RNA per cell, which limits the effectiveness of collecting and processing transcripts prior to sequencing. Also, capture efficacy varies across cells, making direct comparisons difficult. So we used a filter to exclude transcripts with low read counts. This phase guarantees substantial counts, quality checks, and improved performance. Our

proposed method requires a matrix of transcript expression counts, which may be produced using alignment tools like Salmon. Each transcript for a particular gene represents a feature for that gene, and each sample represents observations. We transform read expression data to log scale to show underlying biological processes because natural fold change does not match to the normal distribution, but log2 converted fold change does. Our selected method MANOVA assumes data have a normal distribution, we use the Interquartile Range Rule to remove outliers from control and treatment groups and replace them with the group median. We used the T principle component to predict differentially expressed genes and provide MANOVA p values for the gene's differential expression. The next step is to see whether there is an isoform switch between transcripts of the gene. The median read counts for a group are computed for a transcript. Then the ratios are computed. The median read ratios of paired transcripts were compared. An isoform switch occurs when one transcript of a gene is elevated while another transcript of the same gene is downregulated. We include the kinds of switches (exonic, intronic, 3' UTR, 5' UTR), the fold change difference, and the p value associated with each gene.

1.3 Contributions and Novelties

We propose a method to analyze differential usage of isoforms between control and treatment groups. Our proposed method enables user to identify differentially expressed transcripts that have been overlooked by gene level differential expression analysis. Our contributions are as follows:

- Our proposed method uses raw transcript read counts where other tools uses isoform fraction where each isoform read counts are normalized by total isoform read counts. This method enables user to differentiate switches even there are no change in isoform fraction.
- Our proposed method compares all possible isoform pairs and reports if there is a switch between with log fold change difference and p values and end results are left to end-user to interpret.
- Our proposed method uses median read counts instead of mean read counts.

This step with log transformation prevents outliers from affecting general behaviour and offers robust results. Other tools avaliable generally uses mean read counts where results get easily affected by outliers.

- Our proposed method only considers genes having isoform switch by calculating median read counts before applying MANOVA. This prefiltering mechanism speeds up our process by almost 10 fold.
- Our proposed method offers transcript type information to aid alternative splicing analysis.

1.4 The Outline of the Thesis

The thesis is arranged as follows: Chapter 2 contains all of the terminology and background information. In Chapter 3, the data collection and the proposed framework are discussed in detail. Experiment configuration and experimental findings are reported in Chapter 4 as well as discussion of shortcomings and future improvements. Concluding remarks are included in Chapter 5.

CHAPTER 2

BACKGROUND

In this chapter, we explain general concepts required to understand the how a single gene makes impact for an organism. We start with translation and transcript dynamics, and moving on to alternative splicing and alternative polyadenylation mechanisms in the context of differentially expressed genes and differential transcript usage. We then explain the type of the data we are dealing with and which methods we used to quantify that data. We briefly explain methods to answer similar problem we are dealing with. Then we describe the methods and tools we employed to get.

2.1 Translation - Transcription dynamics

Genes code for proteins, and proteins are responsible for the function of the cell. As a result, the hundreds of genes that are expressed in a certain cell define what that cell is capable of. It is believed that the flow of information from DNA to RNA to protein offers the cell a possible control point for self-regulating its activities by changing the quantity and types of proteins produced by the cell. The quantity of a specific protein present in a cell at any given moment indicates the balance between the protein's synthesis and degradative biochemical pathways at that point in time. To understand the synthetic side of this equation, consider that protein synthesis begins with transcription (DNA to RNA) and continues with translation (RNA to protein). The quantities and kinds of mRNA molecules found in a cell are indicative of the cell's overall activity. When it comes to gene expression, the most important control point is typically found at the very beginning of the protein synthesis process – at the time of transcription. Because a single mRNA molecule may result in the production

of multiple copies of the same protein, RNA transcription is an effective control point [29]. In certain genes, just a portion of the DNA sequence is used in the production of protein. When an RNA transcript (or the DNA that encodes it) has noncoding portions, those sections are called introns, and they are spliced away before the RNA molecule is translated into protein. Exons are the portions of DNA (or RNA) that contain the genetic information for proteins. New, immature strands of messenger RNA, known as pre-mRNA, are produced after transcription and may include both introns and exons. Splicing is a process in which the pre-mRNA molecule undergoes alteration in the nucleus, during which the noncoding introns are removed and only the coding exons are retained. The process of splicing results in the production of a mature messenger RNA molecule, which is subsequently translated into a protein [18].

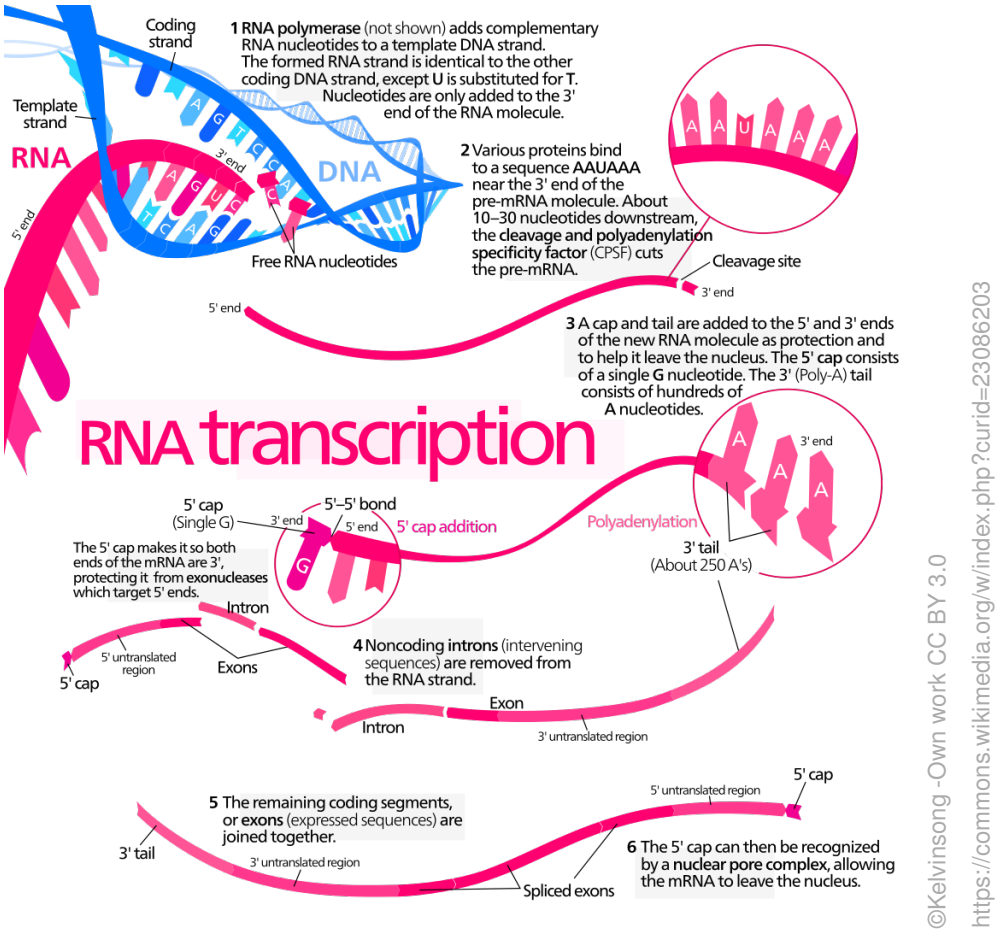

Figure 2.1: Transcription dynamics

The mRNA is made up of a protein-coding region as well as untranslated sections at the 5' and 3' ends (UTRs). The 5' untranslated region (UTR) of an mRNA is the part of the mRNA that extends from the 5' end to the location of the first codon utilized in translation. The segment of an mRNA that extends from the 3' end of an mRNA to the location of the final codon utilized in translation is known as the 3' UTR. The 3' untranslated region (UTR) is variable in both sequence and length; it covers the region between the stop codon and the poly(A) tail. Importantly, the 3' UTR sequence contains numerous regulatory motifs that regulate mRNA turnover, stability, and localization, and, as a result, it is involved in the control of many aspects of post transcriptional gene expression and transcription [35]. In addition, the 5' UTR also contains a large number of binding sites for proteins that either inhibit or enhance translation in response to chemical signals that are relayed. The impairment of any of these characteristics in mRNAs may cause translational regulation to be disrupted, resulting in a variety of illnesses or increased disease vulnerability [7].

2.2 Alternative Splicing

Alternative splicing emerges as a fundamental factor in gene regulation by producing functionally diverse proteomes that interfere with nearly every biological function studied, owing to its widespread use and molecular flexibility [20]. Previous research has revealed that about 92–94 percent of all human genes are affected by alternative splicing, enabling them to express a diverse variety of transcripts from the same genomic region [44]. Such differences result in changes in transcription factors, localization of proteins, enzymatic properties, binding to other proteins, channel proteins and overall changes in cellular proteins [20]. Identification of such transcriptional differences between normal and pathological states may result in the identification of markers of progression or prognosis, as well as the identification of molecules that may be treated with pharmaceuticals or a better understanding of the molecular events that occurred. In most instances, inferring expression changes is a critical but early step of the discovery process, with the findings typically passed into interpretative analysis later [38].

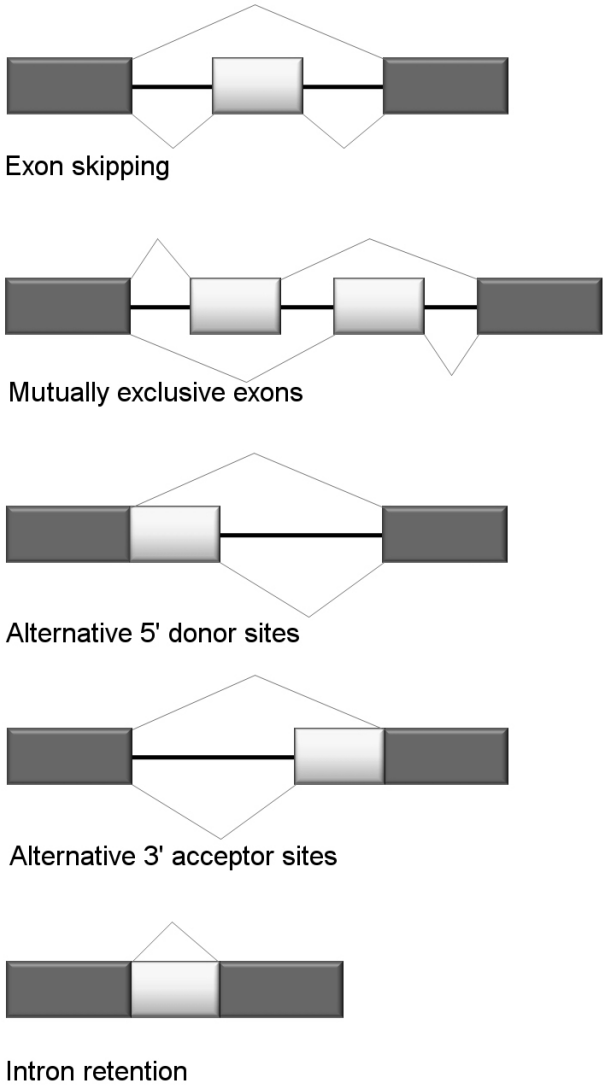

Figure 2.2: basic types of alternative RNA splicing events

2.3 Alternative Polyadenylation

As with alternative splicing, alternative polyadenylation (APA) adds complexity to the human cell transcriptome by generating mRNAs with varied 3' untranslated regions (3'UTRs) and/or encoding variable protein isoforms [13]. Polyadenylation and 3' end formation are essential for gene expression because incorrectly processed messenger RNAs are not transferred from the nucleus to the cytoplasm for translation. Similarly, inefficiently processed mRNAs may be transported out of the nucleus at a

lesser rate than effectively processed mRNAs [25]. By utilizing proximal APA sites and shorter 3'-UTRs, proliferating cells escape miRNA-mediated regulation. These APA and 3'-UTR shortening processes may explain certain proto-oncogene activation in cancer cells [1]. Thus, polyadenylation of 3'UTRs may have a variety of effects on mRNA metabolism, including mRNA stability [46] and translation regulation [5] consequently impacting a cell's development, growth, and viability [25]. Studies have shown that disease-related instances reflecting changes in APA induced by aberrant poly(A) signals exist through loss-of-function mutation or a gain-of-function mutation [10] [17].

2.4 Differentially Expressed Genes

Transcriptome analysis is a valuable technique for identifying and understanding the molecular basis of phenotype diversity in biology, including diseases. The search for differentially expressed (DE) genes is arguably the most prominent use of transcriptome profiling [37]. In order to be classed as differentially expressed, a gene must show a statistically significant difference or change in read counts or expression levels/indexes between two experimental circumstances (e.g., between a control condition and an experimental condition). Gene expression analysis may be accomplished via the use of a number of statistical techniques. Statistical distributions are used to infer the pattern of differential gene expression in order to estimate it. It is common practice to select differentially expressed genes using a combination of an expression change threshold and a score cutoff, which are usually determined via statistical modeling [4].

2.5 Differential Transcript Usage (DTU)

DTU is defined as variations in the expression proportions of a gene's isoforms, as opposed to changes in the amounts of individual transcripts, as previously mentioned. DTU takes part in development and cell differentiation, mRNA repression or stabilization, and in disease mechanisms [45]. For example, cancer cell lines frequently produce significant quantities of mRNA isoforms with shorter 3' untranslated regions

(UTRs) as compared to similarly growing non-transformed cell types [26] which are also related to promotion of tumor growth [30].

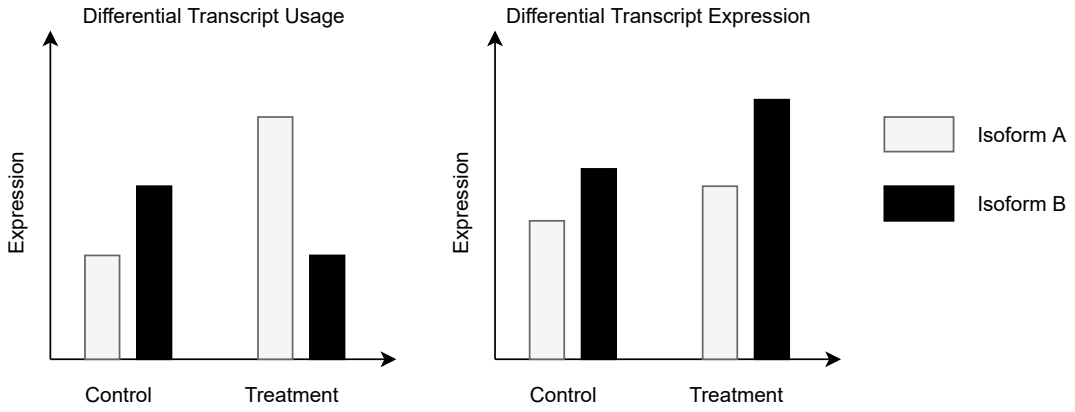

Figure 2.3: For a gene with two isoforms, a schematic representation of differential transcript expression (DTE) and differential transcript usage (DTU) between conditions [38].

2.6 Single Cell RNA-Seq

Individual cell variations can have substantial functional effects on both unicellular and multicellular organisms. The single-cell mRNA sequencing method enables the investigation of individual cells' transcriptomes in an unbiased, high-throughput, and high-resolution manner. Single-cell RNA-sequencing techniques have already revealed new biology in terms of tissue composition, transcription kinetics, and gene regulatory connections [21]. Single-cell data allows us to investigate the regulatory circuitry that regulates cells at a level that was previously impossible to achieve with data gathered from bulk cell populations.

2.7 Salmon

For measuring the amount of transcripts present in cells, next-generation sequencing technology (RNA-seq) has emerged as the standard method. Because of the fragmentation of sequencing material, the measurement and interpretation of RNA-seq data is more complicated [48]. Salmon is a method for predicting transcript abundance from

RNA-seq data that is both fast and reliable to use. When used in conjunction with RNA-sequencing data, Salmon's comprehensive model takes into account sample-specific factors and biases that are common in the data (such as coverage positional errors, sequence-specific biases at the 5' and 3' ends of sequenced fragments), which increases the accuracy of abundance estimations as well as the confidence in subsequent differential expression analyses (e.g., strand-specific procedures and fragment length distribution) [34].

2.8 Related work

When it comes to differential gene expression, there are many tutorials and workflows available. However, the tools available for conducting differential transcript usage (DTU) analysis are scarce. Some of the currently available packages and functions that may be utilized for statistical analysis of DTU are listed below. DEXSeq [2], which was initially developed for differential exon usage identification, makes use of generalized linear models and provides reliable control of false discoveries by taking biological variance into consideration. Differential exon usage may be used as an indication of differential transcript usage, however it is not always possible to determine which isoforms are differentially controlled from one another. Furthermore, it is up to the user to determine whether transcripts are being utilized in a different way. When a single exon of a gene with multiple exons is referred to as differentially utilized, the interpretation of the findings of the technique is simple and clear. When a large number of exons within a gene are affected, the interpretation becomes more difficult. Using the Dirichlet-multinomial distribution as the basis, DRIMSeq [28] offers a statistical framework that may be used to detect variations in isoform usage between circumstances. The Dirichlet-multinomial model naturally accommodates for differential gene expression without sacrificing information about overall gene abundance, and it has the capacity to account for their linked nature by modeling isoform expression along with the other isoform expression models. Sierra [33] uses the differential exon usage technique DEXSeq, but applies it to peak coordinates instead of exons. DTU is detected if there is a substantial shift in the relative utilization of peaks. StageR [11] proposed two-stage testing as a generic paradigm for evaluating

high-throughput studies containing many hypotheses that may be aggregated. When aggregating transcript-level p values to the gene level, the two-stage approach benefits from the impressive performance of FDR control, which is particularly useful for DTU and DTE studies. IsoformSwitchAnalyzeR [42] performs enrichment tests using basic R's prop test, and enrichments are compared using the Fisher test to identify isoform changes across data. In this thesis, we provide end to end analysis starting from transcript level abudances and providing results and illustrations for isoform switches complemented with bio type and transcript type information which may be used to examine a biological phenomenon known as alternative splicing.

2.9 Multivariate Analysis of Variance (MANOVA)

When there are two or more levels in a group, multivariate analysis of variance is used to determine differences between composite means for a collection of dependent variables [41]. In other terms, multivariate analysis of variance (MANOVA) is a variant of analysis of variance (ANOVA) that includes several dependent variables. For example, ANOVA examines the difference in means between two or more groups, whereas MANOVA examines the difference in vectors of meaning between two groups. MANOVA makes several validity assumptions, including multivariate normality, homogeneity of the covariance matrices, and observation independence [16]. The assumption that data follows a multivariate normal distribution is especially implausible for the majority of ecological data sets. This is because the distributions of particular species' abundances are frequently extremely aggregated or skewed. Additionally, abundances are discontinuous rather than continuous, species with tiny means frequently have asymmetric distributions due to being forced to terminate at zero, and uncommon species add a large number of zeros to the data set. MANOVA test statistics are relatively insensitive to deviations from multivariate normality. Finally, many of these test statistics are mathematically difficult to compute when the number of variables exceeds the number of sample units, as is frequently the case in ecological applications [3]. In this context, we use MANOVA to assess statistical significance between read counts of multiple isoforms of a gene from samples with different phenotype.

2.10 Multivariate Independent Comparison of Observations (micompr)

R's package *micompr* is a method for determining if two or more multivariate samples were taken from the same distribution. The technique converts multivariate data to a collection of linearly uncorrelated statistical measures, which are then compared using a variety of statistical methods. This method is insensitive to the distributional characteristics of the samples and automatically chooses the features that best account for their differences. The method is suitable for comparing samples of time series, pictures, spectrometric measurements, or comparable multivariate data with a high dimension. The method works as follows: do a MANOVA test on the samples, assuming that each observation contains q-dimensions matching to the first q principle components (dimensions), and that these dimensions adequately explain a user-defined minimum percentage of variance. Second, use univariate tests to examine observations inside individual PCs. The t-test and the Mann-Whitney U test are two possible tests for comparing two samples, whereas the ANOVA and Kruskal-Wallis tests are the parametric and non-parametric variants for comparing more than two samples, respectively [14].

CHAPTER 3

MATERIALS AND METHODS

3.1 Data acquisition

scRNA-sequencing dataset for primary breast cancer retrieved with GEO Accession number GSE75688. The dataset includes 515 samples from a total of 11 patients having distinct subtypes of breast cancer [9]. Phenotype details can be seen from 3.1. Please refer to the original paper for further library preparation specific information.

			Cell Count					
Patient ID	Subtype	Primary or Metastatic	Bcell	Myeloid	Stromal	Tcell	Tumor	Unknown
BC01	ER+	primary			2		20	4
BC02	ER+	primary					53	3
BC03	ER+ and HER2+	primary	9			9	15	4
BC03LN	ER+ and HER2+	metastatic	38			5	10	2
BC04	HER2+	primary		3	1	4	47	4
BC05	HER2+	primary			1		75	1
BC06	HER2+	primary	8		2		8	7
BC07	TNBC	primary	3	10	4	7	26	1
BC07LN	TNBC	metastatic	23		1	2	26	1
BC08	TNBC	primary		1	6		15	1
BC09	TNBC	primary	1	16	2	7		3
BC09_Re	TNBC	primary	1	6	2	20		2
BC10	TNBC	primary		2	2		11	1
BC11	TNBC	primary					11	

Table 3.1: Cell counts of respective phenotypes

The expression levels of GENCODE v27 genes was estimated using Salmon (version 1.2.0) [34]. Salmon handles alignment to genome, sorting the alignments and quantification of expression levels.

3.2 Proposed method

Here is the general workflow for single cell RNA sequencing analysis and our proposed method visualized in Figure 3.1. After conventional steps like library preparation, alignment to genome , sorting the alignment and expression level quantification, our method takes over and filters lowly expressed transcripts, performs differential expression analysis and identifies genes with isoform switch. After all, isoform switches are reported and data is ready for functional enrichment analysis. We give the details of each individual steps of the proposed method in the following subsections.

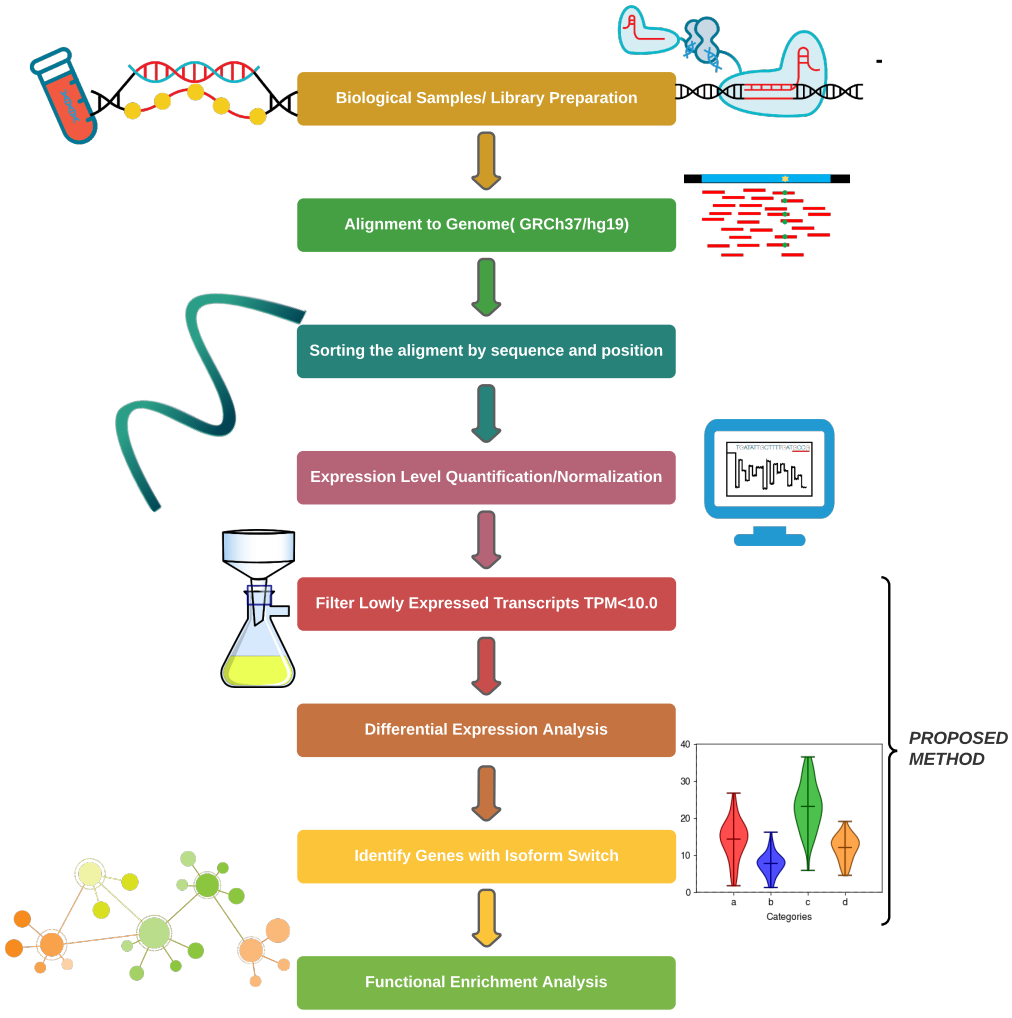

Figure 3.1: The general workflow and proposed method

3.2.1 Filter lowly expressed transcripts

In single cell RNA sequencing, the read counts frequently contain significant amounts of technical noise and a large number of dropouts, i.e., zero or near-zero results. This is owing to the limited quantity of RNA present per cell, which reduces the efficiency with which transcripts can be collected and processed prior to sequencing. Also, capture effectiveness varies from cell to cell, making direct comparisons between cells impossible [24]. Therefore, we employed a filtering mechanism to filter out transcripts that have insignificant read counts. Let n_s be the sample size in the experiment group. This phase requires that the transcripts have a minimum of 10 transcripts per million in at least 10% of the n_s samples. This type of TPM thresholding mechanism is a common method employed in previous studies [12, 23, 33, 38]. This step ensures transcript in consideration have significant counts, have passed quality checks and are also shown to improve performance [38]. Figure 3.2 shows how the filtering threshold rate affects the number of isoform counts analyzed.

3.2.2 Differential Expression Analysis

Our approach needs a matrix of expression counts at the transcript level, which may be generated using alignment tools such as Salmon [34]. Let R_{ts} represent the read expression value for a particular transcript $t = 1, ...T$ in a sample set $s = 1, ..., n$ from a control and treatment group, respectively. The outline of the data can also be expressed like this: each transcript for a specific gene represents a feature for that gene and each sample represents observations. To better highlight the underlying biological processes, read expression data is changed to log scale. Since log fold change is significant in our study because natural fold change does not correspond to the normal distribution, while $log2$ converted fold change does. Since our selected statistical method, MANOVA, assumes data to be in the form of a normal distribution, we prune outliers in control and treatment groups using the Interquartile Range Rule [43] and replace them with the group median. Then, we adopted the Micompr [15] $cmpoutput$ function to model differentially expressed genes with a T principle component. We report MANOVA p values for the significance of differential expression of that gene. Genes with $p<0.05$ are determined as statistically significant.

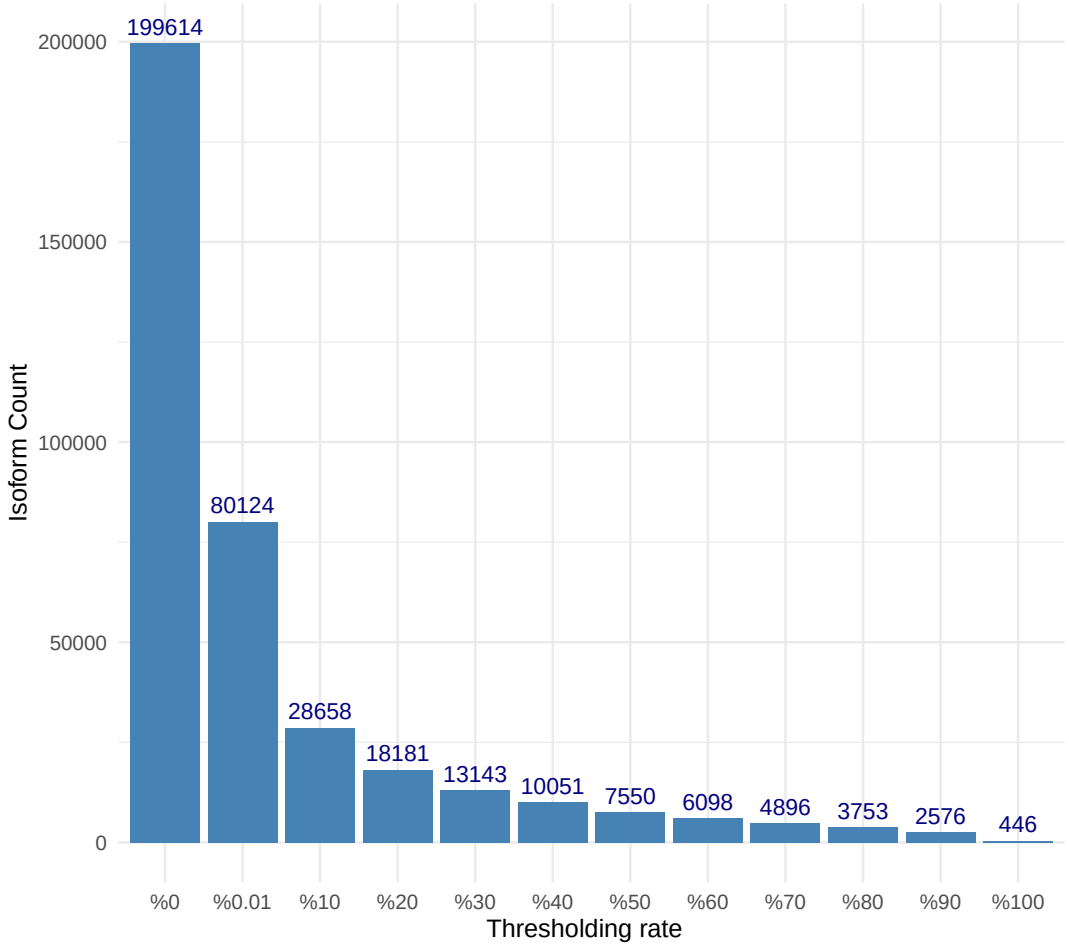

Figure 3.2: Number of transcripts expressing minimum 10 TPMs in respect to threshold rate of sample size

3.2.3 Identify Genes with Isoform Switch

After transcript reads from the gene are deemed to be statistically significant between the control and treatment groups, next step is to find out whether there is an isoform switch between transcripts of the gene. First, transcript reads from samples are grouped according to user-specified phenotypes and then the median read counts for a group are calculated for a transcript. Then, $ratio$ is calculated between groups. Paired combinations of transcripts were matched against each other by comparing their median read ratios. If for a specific transcript of a gene, the control group is up regulated with respect to the treatment group, whereas for a different transcript of the same gene, the control group is down regulated with respect to the treatment group, an

isoform switch is found for the gene. We report pairs of transcripts that exhibit such switches, their types (exonic, intronic, 3' UTR, 5' UTR), their fold change difference and p value associated with that gene.

3.2.4 Transcript Type Assignment

Genomic annotations of human genes and transcripts were downloaded from the GENCODE website(Release 27) [19]. The annotation file is parsed and we retrieved gene related *chromosome, start position (start), end position (end), gene name, strand, coding sequence start position (cds_start), coding sequence end position (cds_end), gene symbol, exon count, list of exon start positions, list of exon end positions* information and then created a hash table for fast access. For transcript information, the annotation file is parsed and we retrieved transcript related *chromosome, strand, transcript start position(start), transcript end position(end), gene id* information and then, we created an iterable list. Then, for each transcript's transcription end point, with respect to strand constraint, compared aganist starting and end points of exon(CDS), intron , 3'UTR, and 5'UTR regions. The pseudo code to get type information also shown in Algorithm 1. Resulting $type$ knowledge joined with isoform switching results gives the user useful information for downstream alternative splicing and APA analysis.

Algorithm 1 Calculate Transcript Type

for each transcript tx from $gene$ **do**
  if $gene[strand]$ is $+$ **then**
    if $tx[end] > gene[start]$ and $tx[end] < gene[cds\_start]$ **then**
      $type \leftarrow 5'UTR$
    end if
    if $tx[end] > gene[cds\_end]$ and $tx[end] < gene[end]$ **then**
      $type \leftarrow 3'UTR$
    end if
    for each exon e of $gene$ **do**
      if $tx[end] \geq e[i][start]$ and $tx[end] \leq e[i][end]$ **then**
        $type \leftarrow Exon\_i$
      end if
      if $tx[end] > e[i][end]$ and $tx[end] < e[i+1][start]$ **then**
        $type \leftarrow Intron\_i$
      end if
    end for
  end if
  if $strand$ is $-$ **then**
    if $tx[start] < gene[end]$ and $tx[start] > gene[cds\_end]$ **then**
      $type \leftarrow 5'UTR$
    end if
    if $tx[start] < gene[cds\_start]$ and $tx[start] > gene[start]$ **then**
      $type \leftarrow 3'UTR$
    end if
    for each exon e of $gene$ **do**
      if $tx[start] \leq e[i][end]$ and $tx[start] \geq e[i][start]$ **then**
        $type \leftarrow Exon\_i$
      end if
      if $tx[start] < e[i][start]$ and $txstart > e[i+1][end]$ **then**
        $type \leftarrow Intron\_i$
      end if
    end for
  end if
end for

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Experiment setup

We evaluated the performance of our proposed method on single-cell RNA-seq data (GEO Accession number GSE75688) [9] containing 549 primary breast cancer cells and lymph node metastases from 11 patients with distinct molecular subtypes (see Figure 3.1) and matched bulk tumors (in our experiments, we did not use any bulk samples and focused only on the single-cell samples). We analyzed differential transcript usage and isoform switches between primer and lymph node metastasis tumor samples of patient BC07 (52 cells in total) because this patient's samples were the only samples that showed consistent transcript levels across samples within a group. Violin plots are generated using the $ggstatsplot$ package [32]

4.2 Filtering

The analysis begins with 199614 transcript read counts from 57876 genes for 52 samples, with 26 samples in each group. Filtering poorly expressed transcripts was performed as stated in 3.2.1, and it resulted in the removal of 86% (n=170956) of all transcripts and 77% (n = 44276) of all genes owing to low expression. The remaining 14% (n = 28658) transcripts relate to the remaining 23% (n= 13600) genes. More than half of these genes (55%, or 7447 of them) express just one transcript and therefore are ineligible to look for isoform switches.

4.3 Isoform switches

The remaining percent 45% were examined for isoform switches, and 1724 transcripts from 626 genes expressed switched read counts between conditions, indicating that isoform switching occurred. After performing MANOVA on 626 genes, we discovered that p values were significant for 414 genes. Among these 414 genes, Bartlett's Test showed that variance differences are negligible for 210 of them, resulting in identical variances for all samples. There were 928 transcript pairs identified for isoform switching in 414 genes where the p values are significant in the context of isoform switching, out of 1724 transcripts to be evaluated. Of these, 482 paired transcript sequences originate from 210 genes, and Bartlett's tests of significance were determined to be negligible. Table 4.1 shows the most significant isoform switches found according to analysis.

	Gene ID	Transcript 1	Type	Biotype	Length	Transcript 2	Type	Biotype	Length	Manova P value	Fold Change Difference
1	ENSG00000013441.15	ENST00000321356.8	Exon_12	protein_coding	484	ENST00000472679.1	3UTR	retained_intron	NA	3.21e-27	2.1955
2	ENSG00000013441.15	ENST00000432425.5	3UTR	nonsense_mediated_decay	NA	ENST00000472679.1	3UTR	retained_intron	NA	3.21e-27	6.9730
3	ENSG00000013441.15	ENST00000472679.1	3UTR	retained_intron	NA	ENST00000409769.6	3UTR	protein_coding	307	3.21e-27	2.7481
4	ENSG00000092199.17	ENST00000336053.10	Exon_7	protein_coding	288	ENST00000553300.5	3UTR	protein_coding	293	1.75e-19	3.7128
5	ENSG00000092199.17	ENST00000556142.5	3UTR	protein_coding	231	ENST00000553300.5	3UTR	protein_coding	293	1.75e-19	7.6719
6	ENSG00000092199.17	ENST00000554455.5	3UTR	protein_coding	306	ENST00000553300.5	3UTR	protein_coding	293	1.75e-19	3.9761
7	ENSG00000092199.17	ENST00000430246.6	3UTR	protein_coding	293	ENST00000553300.5	3UTR	protein_coding	293	1.75e-19	7.2001
8	ENSG00000092199.17	ENST00000553300.5	3UTR	protein_coding	293	ENST00000420743.6	3UTR	protein_coding	306	1.75e-19	8.7423
9	ENSG00000092199.17	ENST00000553300.5	3UTR	protein_coding	293	ENST00000557033.1	5UTR	retained_intron	NA	1.75e-19	6.6742
10	ENSG00000166797.10	ENST00000300030.7	Exon_5	protein_coding	160	ENST00000558779.1	3UTR	processed_transcript	NA	1.05e-17	5.0363
11	ENSG00000166797.10	ENST00000300030.7	Exon_5	protein_coding	160	ENST00000380290.7	3UTR	protein_coding	102	1.05e-17	5.5567
12	ENSG00000112514.15	ENST00000488034.5	3UTR	protein_coding	179	ENST00000607266.5	3UTR	protein_coding	156	1.11e-15	2.9124
13	ENSG00000153207.14	ENST00000470300.5	Exon_36	processed_transcript	NA	ENST00000326225.3	Exon_36	protein_coding	2275	1.77e-15	5.3760
14	ENSG00000041357.15	ENST00000560217.5	3UTR	protein_coding	230	ENST00000413382.6	3UTR	protein_coding	190	5.07e-13	2.8924
15	ENSG00000041357.15	ENST00000044462.11	Exon_9	protein_coding	261	ENST00000413382.6	3UTR	protein_coding	190	5.07e-13	2.8428
16	ENSG00000041357.15	ENST00000559082.5	3UTR	protein_coding	261	ENST00000413382.6	3UTR	protein_coding	190	5.07e-13	2.6696
17	ENSG00000041357.15	ENST00000413382.6	3UTR	protein_coding	190	ENST00000559146.5	Exon_8	protein_coding	NA	5.07e-13	1.5355
18	ENSG00000041357.15	ENST00000413382.6	3UTR	protein_coding	190	ENST00000557929.5	Exon_8	processed_transcript	NA	5.07e-13	2.3999
19	ENSG00000112701.17	ENST00000370010.6	3UTR	protein_coding	1105	ENST00000503501.1	3UTR	nonsense_mediated_decay	NA	5.69e-12	6.6680
20	ENSG00000112701.17	ENST00000436928.7	Exon_9	processed_transcript	NA	ENST00000503501.1	3UTR	nonsense_mediated_decay	NA	5.69e-12	5.1935

Table 4.1: Top 20 isoform switches reported

Figure 4.1 shows that "protein coding - protein coding" that was the most often seen Ensembl transcript biotype pair engaged in isoform switch events, followed by "retained intron (transcripts containing intronic sequences) - protein coding" and "protein coding - processed transcript (transcripts not containing an ORF)".

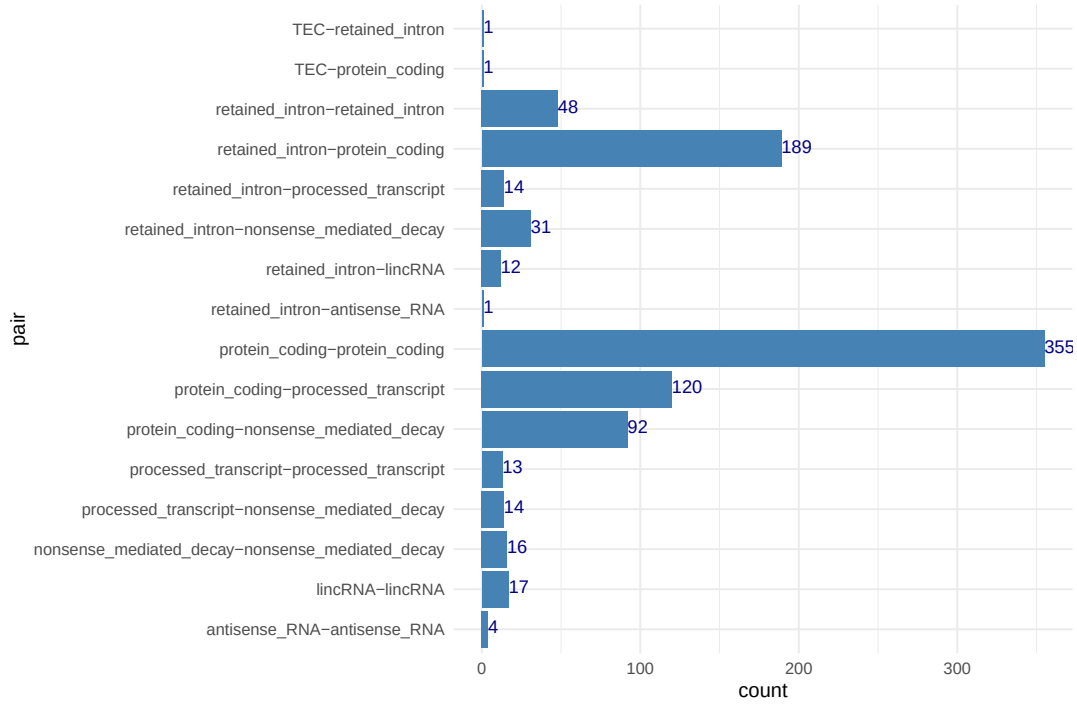

Figure 4.1: Isoform switched transcript pair' biotype distribution

4.4 Biological Interpretation

4.4.1 Isoform switches found for CLK1 gene

The study found that the CLK1 protein regulates alternative RNA splicing through its C-termini and modulates the cellular distribution and splicing activity of the SR family splicing factor, a protein that is very important in pre-mRNA alternative splicing and translation [47]. Increased CLK1 expression was shown to enhance the development and metastasis of pancreatic cancer cells in both vitro and vivo, according to a recently published research [8]. With the help of our analysis, researchers might put light on the underlying transcripts involved in the process.

For gene CLK1 (ENSG00000013441), the following isoform switches were found between:

- protein coding, exonic, ENST00000321356.8 transcript up-regulated in primer tumor samples
- nonsense mediated decay, 3'UTR, ENST00000432425.5 transcript up-regulated

in primer tumor samples, non-existent in metastasis samples

- protein coding, 3'UTR, ENST00000409769.6 transcript up-regulated in primer tumor samples, down-regulated in metastasis samples

- retained intron, 3'UTR, ENST00000472679.1 transcript down regulated in primer tumor samples, up-regulated in metastasis samples

P value reported for the gene is $3.21e^{-27}$ and log fold change difference values are reported as following:

- 2.19 (between $ENST00000321356$ and $ENST00000472679$) (see Figure:4.2)

- 2.74 (between $ENST00000472679$ and $ENST00000409769$) (see Figure:4.3)

- 6.97 (between $ENST00000432425$ and $ENST00000472679$) (see Figure:4.4)

Results from the GTEX database also confirmed that transcripts of CLK1: $ENST00000321356$, $ENST00000432425$, $ENST00000472679$, $ENST00000409769$ are expressed in the breast tissue. Also to assess functional enrichment, we used the STRING database [40]. The analysis of CLK1 showed clusters of enriched pathways related to mRNA 5-splice site recognition, mRNA cis splicing, via spliceosome, spliceosomal complex assembly, regulation of RNA splicing, and regulation of mRNA splicing via spliceosome as the most significant functional enrichments.

In the switch between $ENST00000432425$ and $ENST00000472679$ (see Figure:4.4), metastasis samples show no expression of nonsense mediated decay. Nonsense-mediated RNA decay (NMD) was first identified as a cellular surveillance mechanism in eukaryotic cells that protects the integrity of mRNA transcripts [6]. NMD inhibits translation of mutant mRNAs with premature termination codons (PTCs) by degrading them. Recent research, on the other hand, has shown that NMD plays a considerably wider function in gene expression by controlling the stability of a wide range of normal transcripts [27]. Given that, in metastasis samples, mRNA regulation mechanisms may be affected due to lack of NMD. The up-regulation of retained intron $ENST00000472679$ can be explained by this affected regulation mechanism.

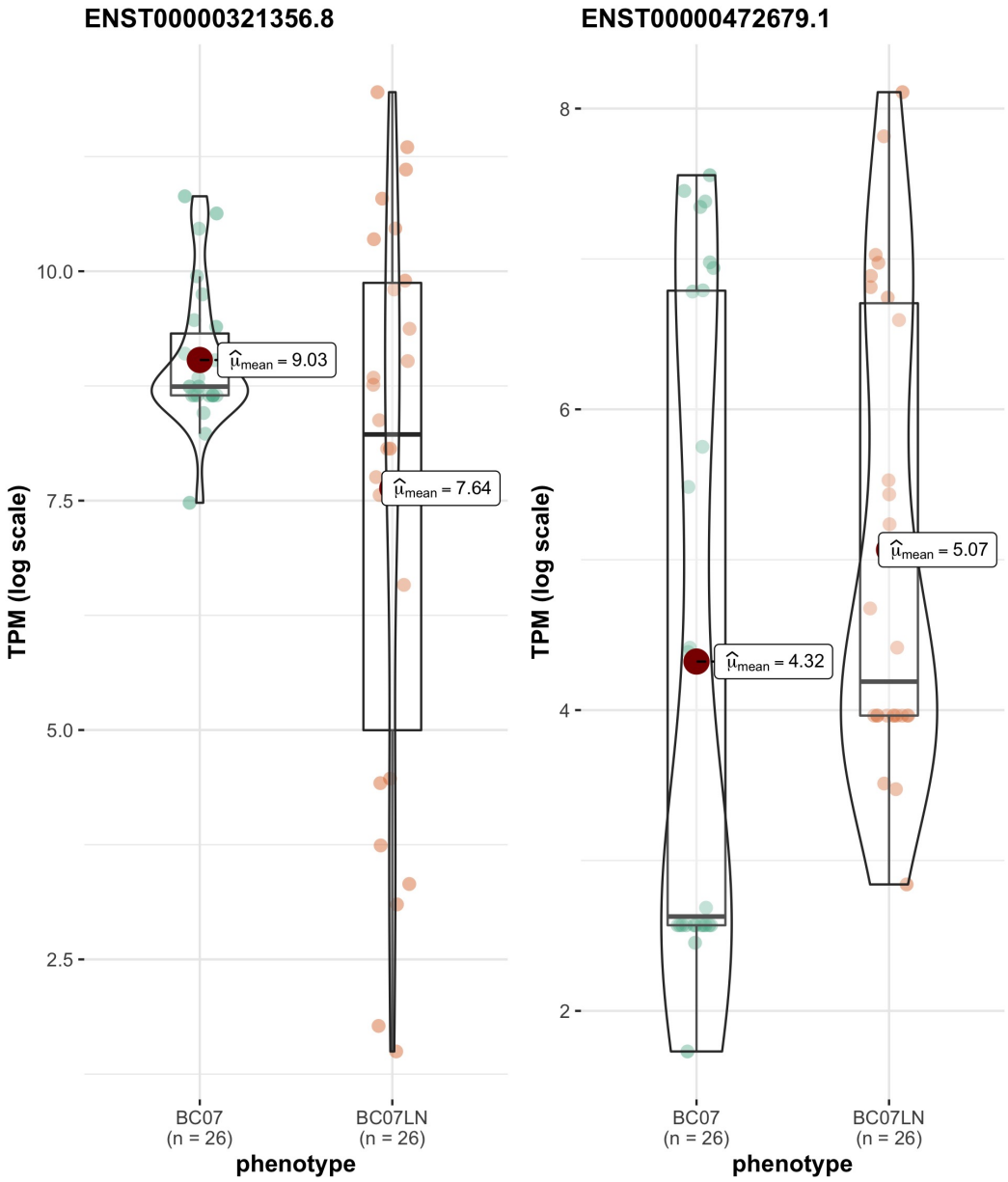

Figure 4.2: Violin plot for isoform switching case CLK1(ENSG00000013441) between protein coding ENST00000321356(Exonic) and retained intron ENST00000472679(3'UTR)

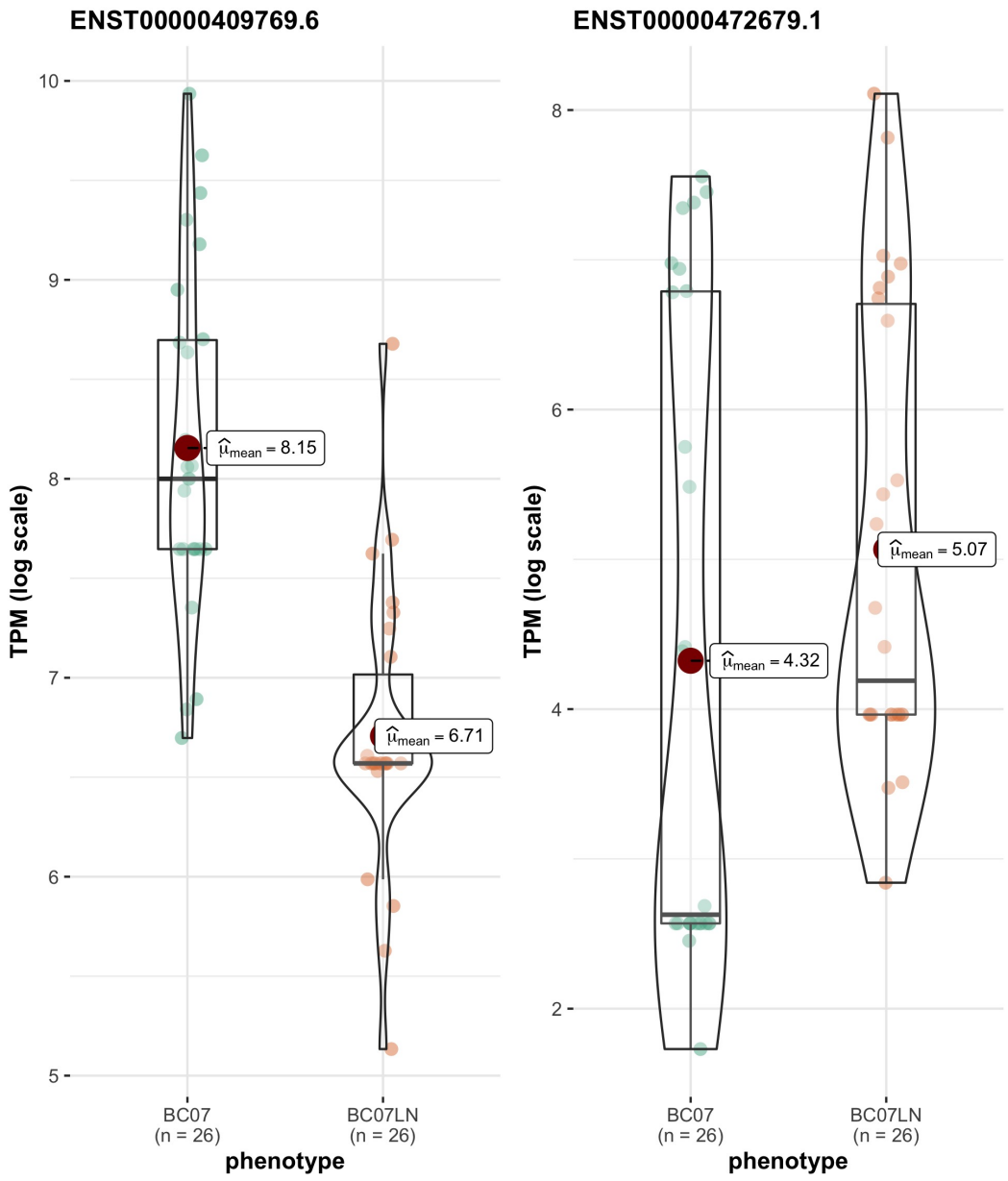

Figure 4.3: Violin plot for isoform switching case CLK1(ENSG00000013441) between protein coding ENST00000409769(3'UTR) and retained intron ENST00000472679(3'UTR)

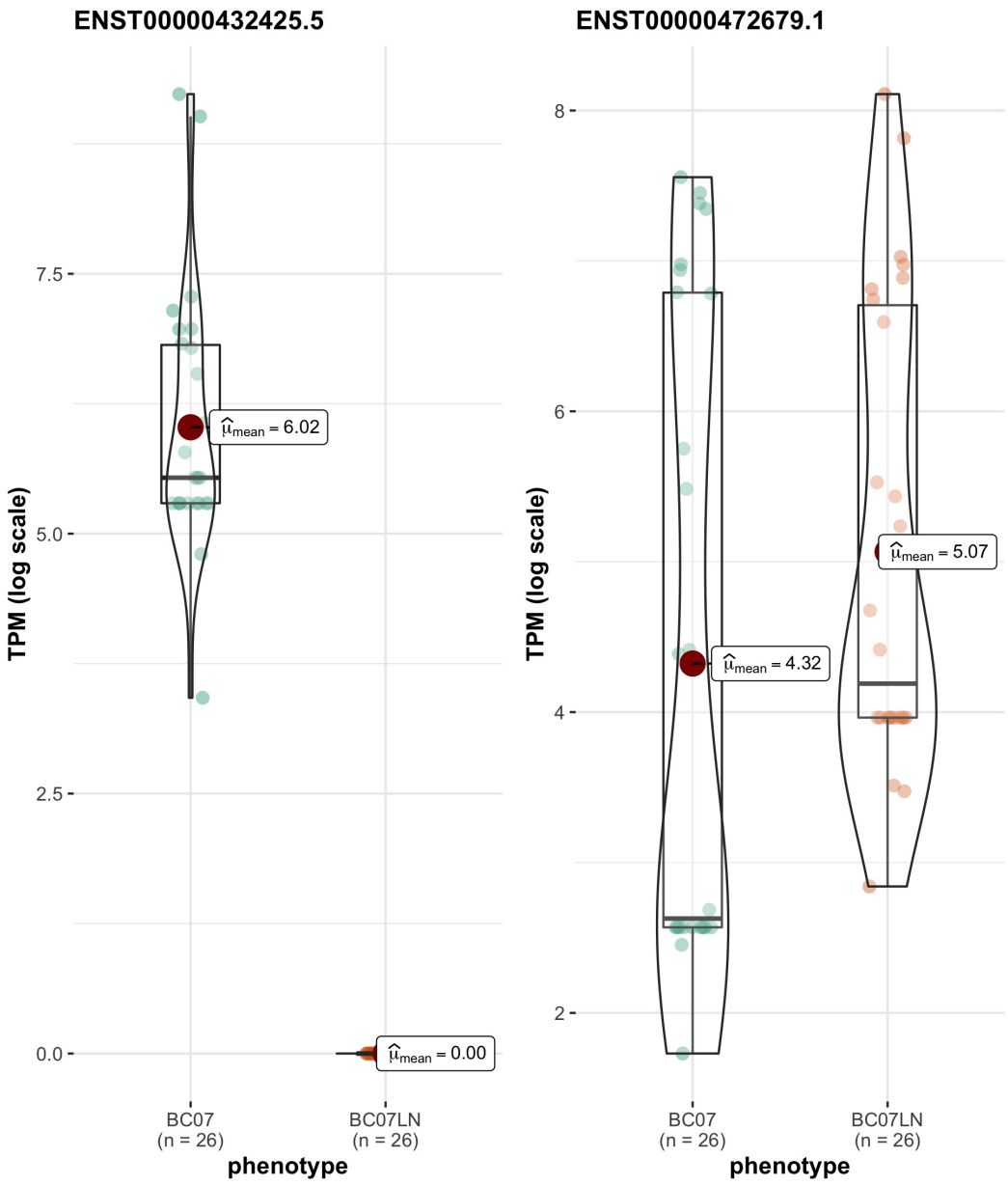

Figure 4.4: Violin plot for isoform switching case CLK1(ENSG00000013441) between nonsense mediated decay ENST00000432425(3'UTR) and retained intron ENST00000472679(3'UTR)

4.4.2 Isoform switches found for HNRNPC gene

HNRNPC is related to 3'-UTR-mediated mRNA stabilization [36], mRNA splicing [22] [39]. Recent research has shown that HNRNPC regulates the aggressiveness of tumor cells in the brain via regulating the expression of the protein PDCD4 (Programmed cell death 4). microRNA-21 levels were decreased as a result of hnRNPC silencing, which resulted in increased PDCD4 expression and a reduction in migratory and invasive behaviors. This highlights the potential use of hnRNPC as a predictive and therapeutic marker in the treatment of cancer [31].
For gene HNRNPC (ENSG00000092199), the following isoform switches were found between:

- protein coding, exonic, ENST00000336053 transcript up-regulated in primer tumor samples, down-regulated in metastasis samples
- protein coding, 3'UTR, ENST00000556142 transcript up-regulated in primer tumor samples, down-regulated in metastasis samples
- protein coding, 3'UTR, ENST00000554455 transcript up-regulated in primer tumor samples, down-regulated in metastasis samples
- protein coding, 3'UTR, ENST00000430246 transcript up-regulated in primer tumor samples, down-regulated in metastasis samples
- protein coding, 3'UTR, ENST00000553300 transcript down-regulated in primer tumor samples, up-regulated in metastasis samples
- protein coding, 3'UTR, ENST00000420743 transcript up-regulated in primer tumor samples, down-regulated in metastasis samples
- retained intron, 5'UTR, ENST00000557033 transcript up-regulated in primer tumor samples, down-regulated in metastasis samples

P value reported for the gene is $1.75e^{-19}$ and log fold change difference values are reported as following:

- 3.71 (between $ENST00000336053$ and $ENST00000553300$) (see Figure:4.5)

- 7.67 (between $ENST00000556142$ and $ENST00000553300$) (see Figure:4.6)

- 3.97 (between $ENST00000554455$ and $ENST00000553300$) (see Figure:4.7)

- 7.20 (between $ENST00000430246$ and $ENST00000553300$) (see Figure:4.8)

- 8.74 (between $ENST00000553300$ and $ENST00000420743$) (see Figure:4.9)

- 6.67 (between $ENST00000553300$ and $ENST00000557033$) (see Figure:4.10)

Results from the GTEX database also confirmed that transcripts of HNRNPC: $ENST00000336053$, $ENST00000556142$, $ENST00000554455$, $ENST00000430246$, $ENST00000553300$ and $ENST00000557033$ are expressed in the breast tissue. Isoform switches are reported mostly between protein coding transcripts. These switches may result in functional gain or loss of function depending on the individual protein traits. Down-regulation of heterogeneous nuclear ribonucleoproteins C1/C2 coding $ENST00000553300$, may lead to decreased regulation performance, resulting in up-regulated retained intron transcripts like $ENST00000420743$. The functional enrichment analysis of HNRNPC showed clusters of enriched pathways related to negative regulation of mRNA splicing, via spliceosome, 3-UTR-mediated mRNA stabilization, negative regulation of telomere maintenance via telomerase, negative regulation of mRNA metabolic process, and the fibroblast growth factor receptor signaling pathway as the most significant functional enrichments.

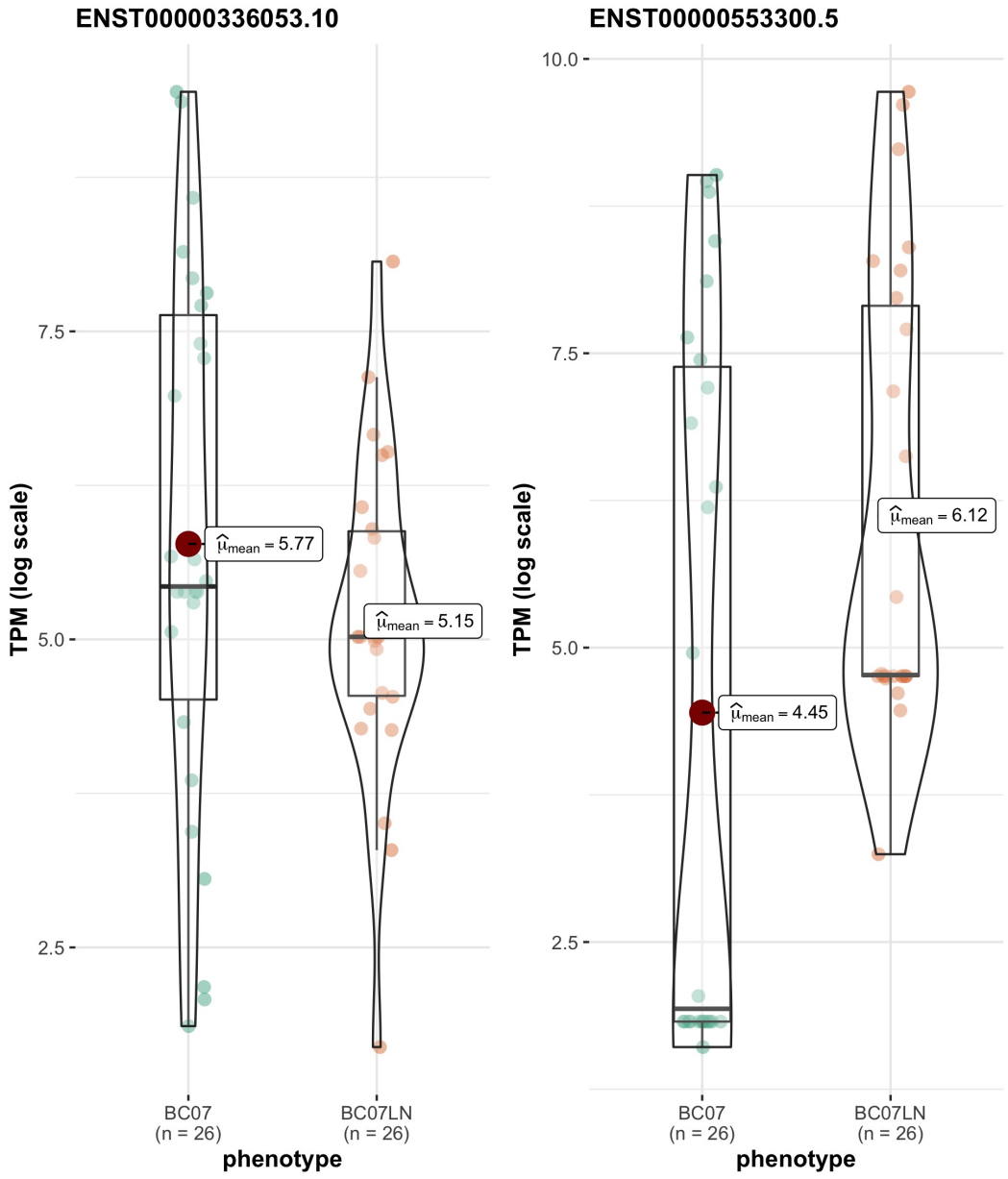

Figure 4.5: Violin plot for isoform switching case HNRNPC(ENSG00000092199) between protein coding ENST00000336053(Exonic) and protein coding ENST00000553300(3'UTR)

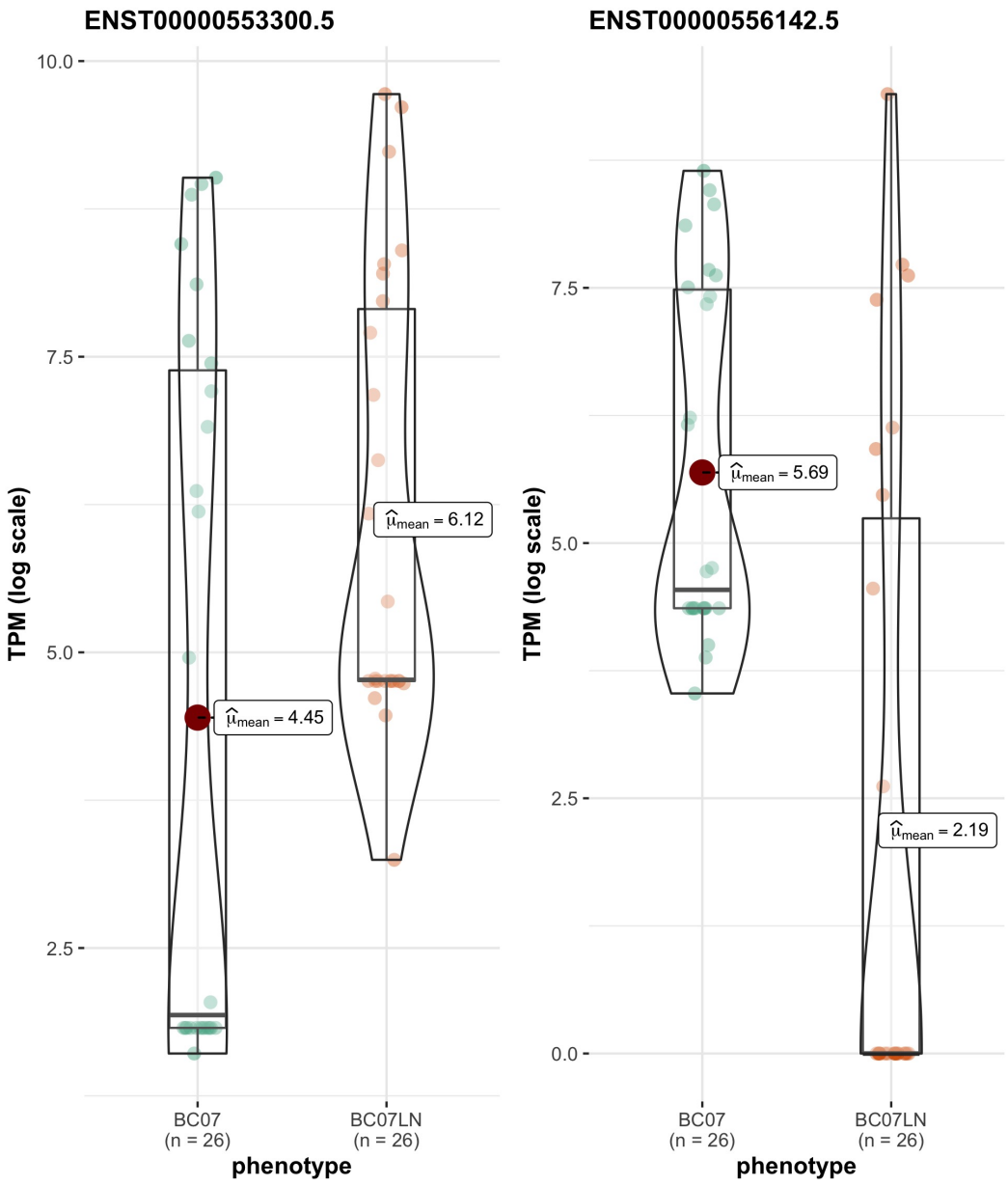

Figure 4.6: Violin plot for isoform switching case HNRNPC(ENSG00000092199) between protein coding ENST00000553300(3'UTR) and protein coding ENST00000556142(3'UTR)

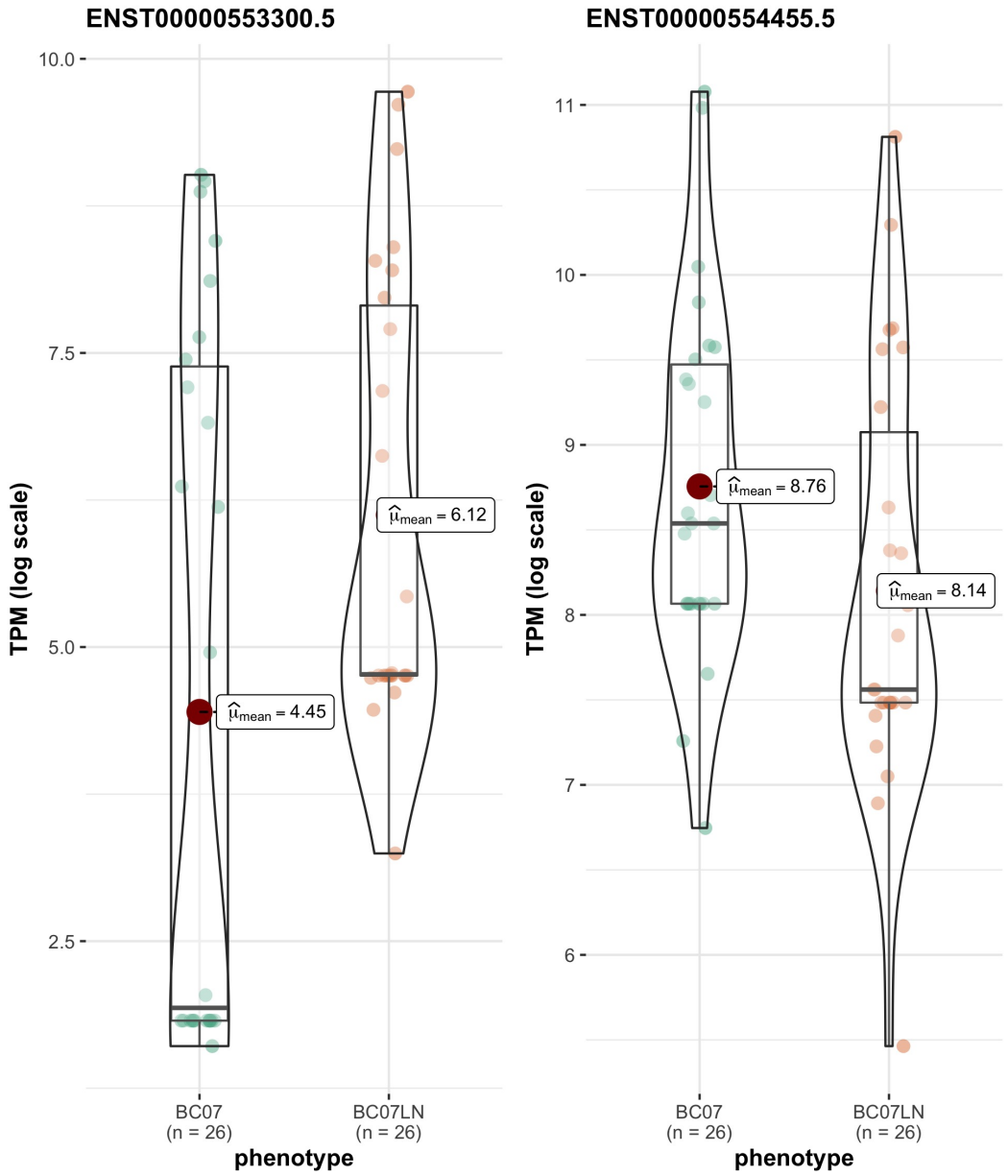

Figure 4.7: Violin plot for isoform switching case HNRNPC(ENSG00000092199) between protein coding ENST00000553300(3'UTR) and protein coding ENST00000554455(3'UTR)

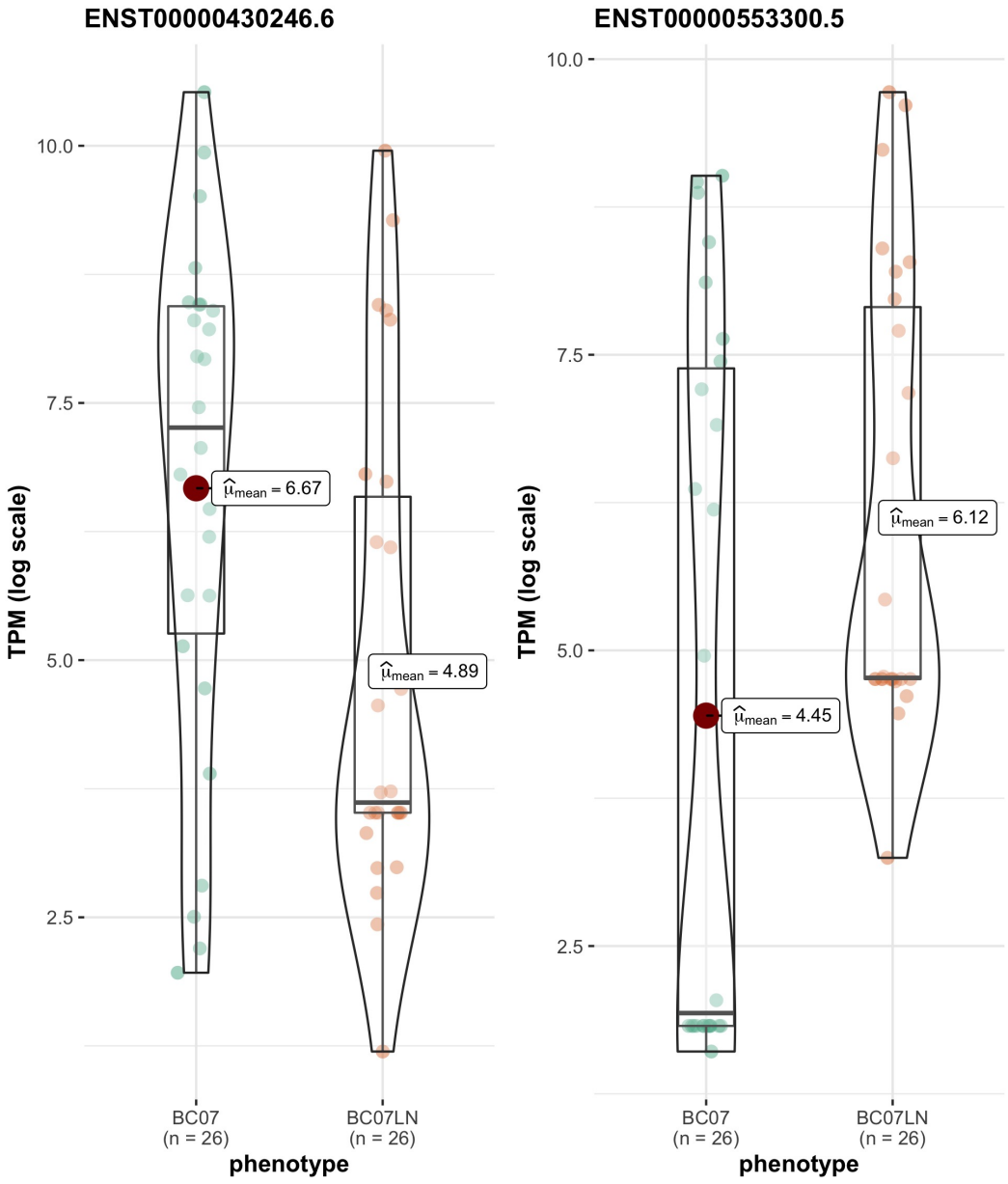

Figure 4.8: Violin plot for isoform switching case HNRNPC(ENSG00000092199) between protein coding ENST00000430246(3'UTR) and protein coding ENST00000553300(3'UTR)

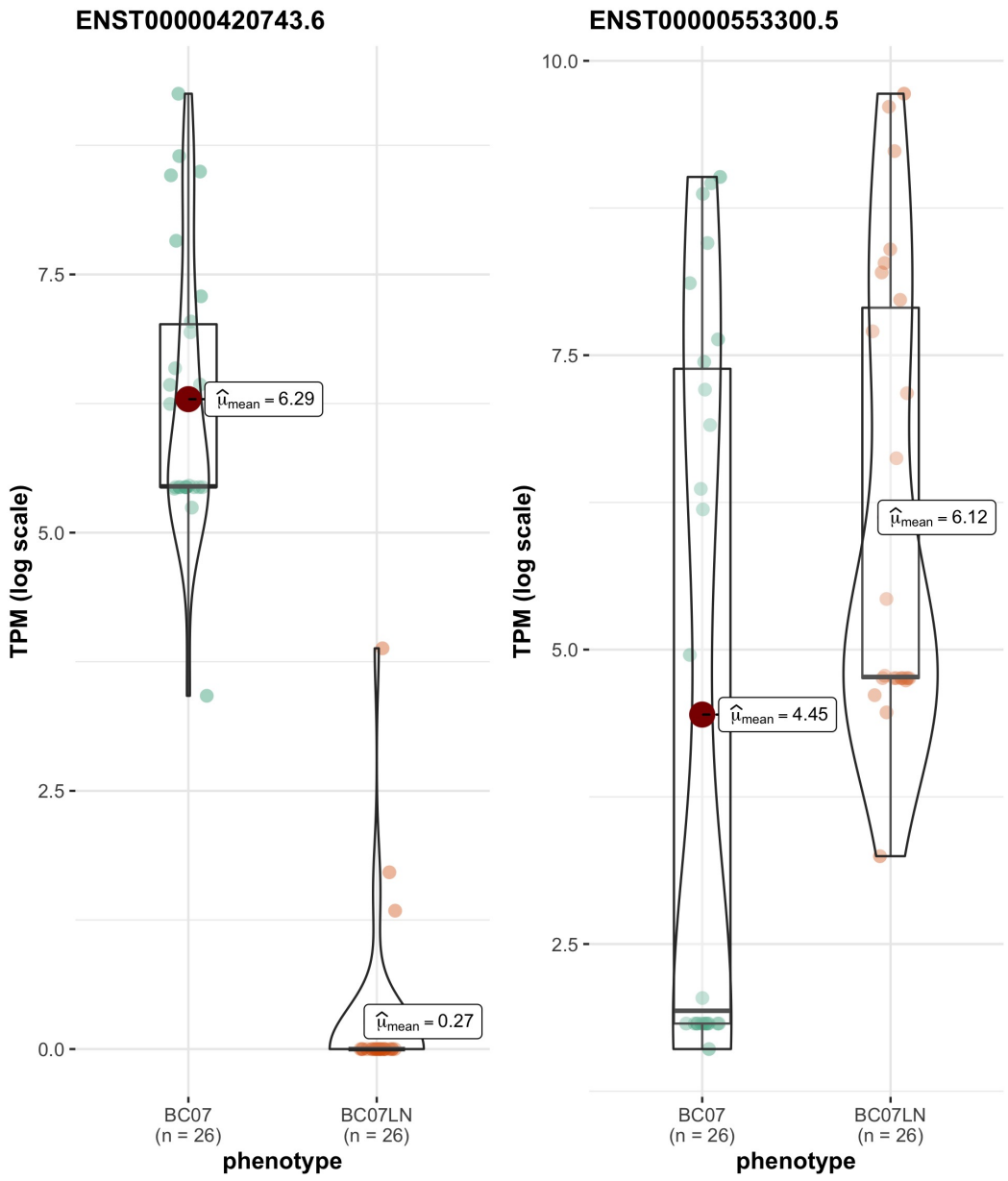

Figure 4.9: Violin plot for isoform switching case HNRNPC(ENSG00000092199) between protein coding ENST00000420743(3'UTR) and protein coding ENST00000553300(3'UTR)

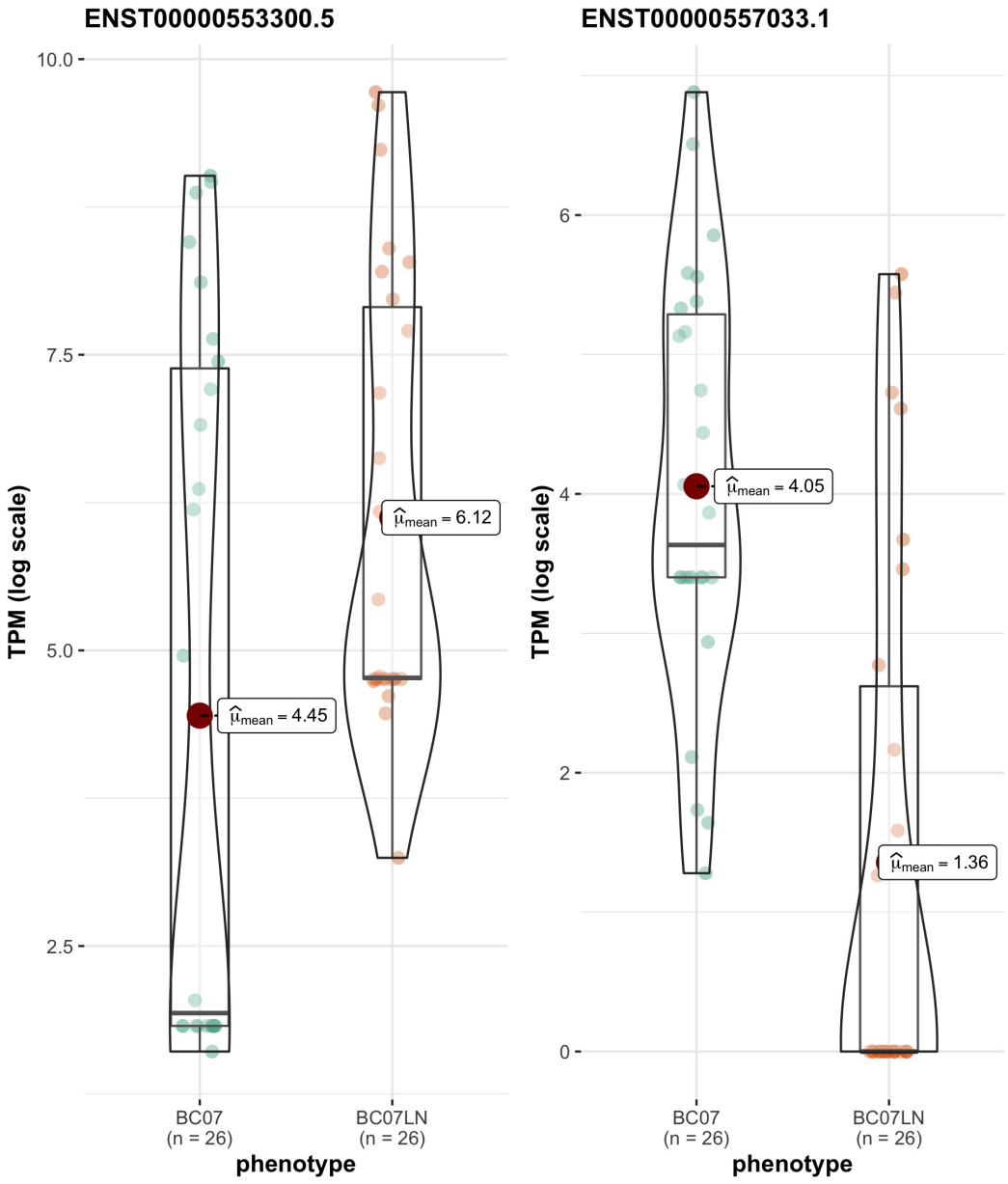

Figure 4.10: Violin plot for isoform switching case HNRNPC(ENSG00000092199) between protein coding ENST00000553300(3'UTR) and protein coding ENST00000557033(5'UTR)

4.5 Discussion

Our method differs from other avaliable methods in following manners:

- Our proposed method uses raw transcript read counts where other tools uses isoform fraction where each isoform read counts are normalized by total isoform read counts. Tools that uses isoform fraction generally requires change in isoform fraction to be higher than predefined value and this value cannot be set by user.
- Our proposed method compares all possible isoform pairs and reports if there is a switch between with log fold change difference and p values and end results are left to end-user to interpret.
- Our proposed method uses median read counts instead of mean read counts. This step with log transformation prevents outliers from affecting general behaviour and offers robust results. Other tools avaliable generally uses mean read counts where results get easily affected by outliers.
- Our proposed method only considers genes having isoform switch by calculating median read counts before applying MANOVA. This prefiltering mechanism speeds up our process by 10 fold through reducing genes to look at from 6911 to 626.
- Our proposed method offers transcript type information to aid alternative splicing analysis.

In all DTU techniques, there is a similar issue in that the ability to identify differentially utilized transcripts is highly dependent on the quality of the corresponding scRNA-seq dataset. This is due to the nature of the sequencing method, which results in a much higher percentage of zero counts in the transcript-level count matrix than in datasets generated by other sequencing methods. Our statistical method, MANOVA, is also sensitive to the non-normality of the dataset as well as variations in sample size. Another disadvantage of MANOVA is that it is unable to build a model when the sample size is limited in comparison to the number of isoforms.

CHAPTER 5

CONCLUSION

We have presented a complete tool that takes transcript read counts and user-specified phenotype grouping as input and provides an extensive isoform switching report (including statistical significance and types of isoform switches) with illustrations as output. Our approach allows for the investigation of changes in genome-wide patterns of alternative splicing and isoform switch alterations using multiple analysis of variance (MANOVA). We have leveraged and identified heterogeneity of isoforms that would have been missed by conventional differential gene expression analysis frameworks. We have analyzed single cell breast cancer data and demonstrated the methods' capability and identified and illustrated several particularly enriched switches. The analysis also includes bio type and transcript type information on top of switching information for each transcript. This metadata information can be used to investigate a biological phenomenon known as alternative splicing, which is currently poorly understood and whose investigation may yield many better treatment and prognosis options. Also, data taken at different stages of disease might shed light on the relative expression levels of transcripts, so the stage of disease can be identified. Transcript based inhibition or promoting mechanisms might lead to novel treatment options. In the future, alternative statistical methods, such as generalized linear models, might be used to explain underlying biological processes or overcome the sparsity of single cell RNA-Seq data. Different types of outlier identification algorithms, in addition to the Interquartile Range Rule, may be offered as choices in the future.

REFERENCES

[1] B. H. Akman, T. Can, and A. E. Erson-Bensan. Estrogen-induced upregulation and 3′-UTR shortening of CDC6. *Nucleic Acids Research*, 40(21):10679–10688, Sept. 2012.

[2] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017, June 2012.

[3] M. J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, Feb. 2001.

[4] A. Anjum, S. Jaggi, E. Varghese, S. Lall, A. Bhowmik, and A. Rai. Identification of differentially expressed genes in RNA-seq data of arabidopsis thaliana: A compound distribution approach. *Journal of Computational Biology*, 23(4):239–247, Apr. 2016.

[5] D. P. Bartel. MicroRNAs. *Cell*, 116(2):281–297, Jan. 2004.

[6] P. Belgrader, J. Cheng, X. Zhou, L. S. Stephenson, and L. E. Maquat. Mammalian nonsense codons can be cis effectors of nuclear mRNA half-life. *Molecular and Cellular Biology*, 14(12):8219–8228, Dec. 1994.

[7] S. Chatterjee and J. K. Pal. Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. *Biology of the Cell*, 101(5):251–262, May 2009.

[8] S. Chen, C. Yang, Z.-W. Wang, J.-F. Hu, J.-J. Pan, C.-Y. Liao, J.-Q. Zhang, J.-Z. Chen, Y. Huang, L. Huang, Q. Zhan, Y.-F. Tian, B.-Y. Shen, and Y.-D. Wang. CLK1/SRSF5 pathway induces aberrant exon skipping of METTL14 and cyclin l2 and promotes growth and metastasis of pancreatic cancer. *Journal of Hematology & Oncology*, 14(1), Apr. 2021.

[9] W. Chung, H. H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, Z. Kan, W. Han, and W.-Y. Park. Single-cell

RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications*, 8(1), May 2017.

[10] B. Conne, A. Stutz, and J.-D. Vassalli. The 3′ untranslated region of messenger RNA: A molecular 'hotspot' for pathology? *Nature Medicine*, 6(6):637–641, June 2000.

[11] K. V. den Berge, C. Soneson, M. D. Robinson, and L. Clement. stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biology*, 18(1), Aug. 2017.

[12] F. Dick, G. S. Nido, G. W. Alves, O.-B. Tysnes, G. H. Nilsen, C. Dölle, and C. Tzoulis. Differential transcript usage in the parkinson's disease brain. *PLOS Genetics*, 16(11):e1009182, Nov. 2020.

[13] G. Edwalds-Gilbert. Alternative poly(a) site selection in complex transcription units: means to an end? *Nucleic Acids Research*, 25(13):2547–2561, July 1997.

[14] N. Fachada, J. ao Rodrigues, V. V. Lopes, R. C. Martins, and A. C. Rosa. micompr: An r package for multivariate independent comparison of observations. *The R Journal*, 8(2):405–420, 2016.

[15] N. Fachada, J. Rodrigues, V. L. Vitor, C. M. Rui, and C. R. Agostinho. micompr: An r package for multivariate independent comparison of observations. *The R Journal*, 8(2):405, 2016.

[16] A. French, M. Macedo, J. Poulsen, T. Waterson, and A. Yu. Multivariate analysis of variance (manova), 2008.

[17] D. C. D. Giammartino, K. Nishida, and J. L. Manley. Mechanisms and consequences of alternative polyadenylation. *Molecular Cell*, 43(6):853–866, Sept. 2011.

[18] W. Gilbert. Why genes in pieces? *Nature*, 271(5645):501–501, Feb. 1978.

[19] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saun-

ders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 22(9):1760–1774, Sept. 2012.

[20] O. Kelemen, P. Convertini, Z. Zhang, Y. Wen, M. Shen, M. Falaleeva, and S. Stamm. Function of alternative splicing. *Gene*, 514(1):1–30, Feb. 2013.

[21] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann. The technology and biology of single-cell RNA sequencing. *Molecular Cell*, 58(4):610–620, May 2015.

[22] N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan. N6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature*, 518(7540):560–564, Feb. 2015.

[23] M. I. Love, C. Soneson, and R. Patro. Swimming downstream: statistical analysis of differential transcript usage following salmon quantification. *F1000Research*, 7:952, Oct. 2018.

[24] A. T. L. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1), Apr. 2016.

[25] C. S. Lutz. Alternative polyadenylation: A twist on mRNA 3′ end formation. *ACS Chemical Biology*, 3(10):609–617, Sept. 2008.

[26] C. Mayr and D. P. Bartel. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4):673–684, Aug. 2009.

[27] A. Nickless, J. M. Bailis, and Z. You. Control of gene expression through the nonsense-mediated RNA decay pathway. *Cell & Bioscience*, 7(1), May 2017.

[28] M. Nowicka and M. D. Robinson. DRIMSeq: a dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, 5:1356, Dec. 2016.

[29] C. M. O'Connor and J. U. Adams. *Essentials of Cell Biology*. MA: NPG Education, Cambridge, 2010.

[30] H. J. Park, P. Ji, S. Kim, Z. Xia, B. Rodriguez, L. Li, J. Su, K. Chen, C. P. Masamha, D. Baillat, C. R. Fontes-Garfias, A.-B. Shyu, J. R. Neilson, E. J. Wagner, and W. Li. 3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. *Nature Genetics*, 50(6):783–789, May 2018.

[31] Y. M. Park, S. J. Hwang, K. Masuda, K.-M. Choi, M.-R. Jeong, D.-H. Nam, M. Gorospe, and H. H. Kim. Heterogeneous nuclear ribonucleoprotein c1/c2 controls the metastatic potential of glioblastoma by regulating PDCD4. *Molecular and Cellular Biology*, 32(20):4237–4244, Oct. 2012.

[32] I. Patil. Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software*, 6(61):3167, 2021.

[33] R. Patrick, D. T. Humphreys, V. Janbandhu, A. Oshlack, J. W. Ho, R. P. Harvey, and K. K. Lo. Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome Biology*, 21(1), July 2020.

[34] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, Mar. 2017.

[35] J. Schwerk and R. Savan. Translating the untranslated region. *The Journal of Immunology*, 195(7):2963–2971, Sept. 2015.

[36] S. Shetty. Regulation of urokinase receptor mRNA stability by hnRNP c in lung epithelial cells. *Molecular and Cellular Biochemistry*, 272(1-2):107–118, Apr. 2005.

[37] C. Soneson and M. Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1), Mar. 2013.

[38] C. Soneson, K. L. Matthes, M. Nowicka, C. W. Law, and M. D. Robinson. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biology*, 17(1), Jan. 2016.

[39] M. S. Swanson, T. Y. Nakagawa, K. LeVan, and G. Dreyfuss. Primary structure of human nuclear ribonucleoprotein particle c proteins: conservation of sequence and domain structures in heterogeneous nuclear RNA, mRNA, and pre-rRNA-binding proteins. *Molecular and Cellular Biology*, 7(5):1731–1739, May 1987.

[40] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, Oct. 2014.

[41] B. G. Tabachnick and L. S. Fidell. *Using multivariate statistics*. Pearson Education, 2013.

[42] K. Vitting-Seerup and A. Sandelin. IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics*, 35(21):4469–4471, Apr. 2019.

[43] X. Wan, W. Wang, J. Liu, and T. Tong. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Medical Research Methodology*, 14(1), Dec. 2014.

[44] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov. 2008.

[45] E. T. Wang, A. J. Ward, J. M. Cherone, J. Giudice, T. T. Wang, D. J. Treacy, N. J. Lambert, P. Freese, T. Saxena, T. A. Cooper, and C. B. Burge. Antagonistic regulation of mRNA expression and splicing by CELF and MBNL proteins. *Genome Research*, 25(6):858–871, Apr. 2015.

[46] C. J. Wilusz and J. Wilusz. Bringing the role of mRNA decay in the control of gene expression into focus. *Trends in Genetics*, 20(10):491–497, Oct. 2004.

[47] S. Yang, R. Jia, and Z. Bian. SRSF5 functions as a novel oncogenic splicing factor and is upregulated by oncogene SRSF3 in oral squamous cell car-

cinoma. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1865(9):1161–1172, Sept. 2018.

[48] S. Zhao, W.-P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu. Comparison of RNA-seq and microarray in transcriptome profiling of activated t cells. *PLoS ONE*, 9(1):e78644, Jan. 2014.